Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification

François Bachoc
Josselin Garnier
Jean-Marc Martinez

CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
LPMA, Université Paris 7

June 2012

The PhD

Two components of the PhD

- Use of Kriging model for code validation
  - Bachoc F, Bois G, and Martinez J.M, Gaussian process computer model validation method, *Submitted*.
- Work on the problem of the covariance function estimation
  - Bachoc F, Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Submitted*.

Cross Validation (Leave-One-Out)

Gaussian Process Y observed at $x_1, ..., x_n$ with values $y = (y_1, ..., y_n)^t$

Cross Validation (Leave-One-Out) principle

- $\hat{y}_{i,-i} = \mathbb{E}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$
- $c^2_{i,-i} = \mathbb{E}((Y(x_i) - \hat{y}_{i,-i})^2|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$

Can be used for Kriging verification or for covariance function selection

Virtual Cross Validation

When mean of Y is parametric : $\mathbb{E}(Y(x)) = \sum_{i=1}^{p} \beta_i h_i(x)$. Let

- $\mathbf{H}$ the $n \times p$ matrix with $\mathbf{H}_{i,j} = h_j(x_i)$
- $\mathbf{R}$ the covariance matrix of $y = (y_1, ..., y_n)$

Virtual Leave-One-Out

With

$$\mathbf{Q}^- = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}$$

We have :

$$y_i - \hat{y}_{i,-i} = (diag(\mathbf{Q}^-))^{-1}\mathbf{Q}^- y \quad \text{and} \quad c_{i,-i}^2 = \frac{1}{(\mathbf{Q}^-)_{i,i}}$$

If Bayesian case for β ($\beta \sim \mathcal{N}(\beta_{prior}, \mathbf{Q}_{prior})$), then same formula holds replacing $\mathbf{Q}^-$ with $(\mathbf{R} + \mathbf{H}\mathbf{Q}_{prior}\mathbf{H}^t)^{-1}$

O. Dubrule, Cross Validation of Kriging in a Unique Neighborhood, *Mathematical Geology*, 1983.

Cross Validation for covariance function estimation (1/2)

Let $\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\}$ be a set of covariance function for Y, with K_θ a correlation function. Let

- $\hat{y}_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$
- $\sigma^2 c^2_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}((Y(x_i) - \hat{y}_{\theta,i,-i})^2|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$

Leave-One-Out criteria we study

$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_{\theta,i,-i})^2$$

and

$$\hat{\sigma}^2_{CV} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{\hat{\theta}_{CV} i,-i})^2}{c^2_{\hat{\theta}_{CV},i,-i}}$$

Cross Validation for covariance function estimation (2/2)

- Leave-One-Out estimation is tractable
- Other Cross-Validation criteria exist

  C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, *The MIT Press, Cambridge*, 2006.
- To the best of our knowledge : problems of the choice of the cross validation criterion and of the cross validation procedure are not fully solved for Kriging
- It is our intuition that when one is primarily interested in prediction mean square error and point-wise estimation of the prediction mean square error, the Leave-One-Out criteria presented are reasonable

Objectives

We want to study the cases of model misspecification, that is to say the cases when the true covariance function K_1 of Y is far from $\mathcal{K} = \left\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\right\}$

In this context we want to compare Leave-One-Out and Maximum Likelihood estimators from the point of view of prediction mean square error and point-wise estimation of the prediction mean square error

We proceed in two steps

- When $\mathcal{K} = \left\{\sigma^2 K_2, \sigma^2 \geq 0\right\}$, with K_2 a correlation function, and K_1 is the true covariance function : Theoretical formula and numerical tests
- In the general case : Numerical studies

Setting

Let x_0 be a new point and assume the mean of Y is zero and K_1 is unit-variance stationary. Let

- r_1 be the covariance vector between $x_1, ..., x_n$ and x_0 with covariance function K_1
- r_2 be the covariance vector between $x_1, ..., x_n$ and x_0 with covariance function K_2
- $\mathbf{R}_1$ be the covariance matrix of $x_1, ..., x_n$.with covariance function K_1
- $\mathbf{R}_2$ be the covariance matrix of $x_1, ..., x_n$.with covariance function K_2

$\hat{y}_0 = r_2^t \mathbf{R}_2^{-1} y$ is the Kriging prediction

$\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] = (r_1^t \mathbf{R}_1^{-1} y - r_2^t \mathbf{R}_2^{-1} y)^2 + 1 - r_1^t \mathbf{R}_1^{-1} r_1$ is the conditional mean square error of the non-optimal prediction

One estimates σ^2 with $\hat{\sigma}^2$ and estimates the conditional mean square error with $\hat{\sigma}^2 c_{x_0}^2$ with $c_{x_0}^2 := 1 - r_2^t \mathbf{R}_2^{-1} r_2$

The Risk

The Risk

We study the Risk criterion for an estimator $\hat{\sigma}^2$ of σ^2

$$R_{\hat{\sigma}^2, x_0} = \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_0}^2\right)^2\right]$$

Formula for quadratic estimators

When $\hat{\sigma}^2 = y^t \mathbf{M} y$, we have

$$\begin{aligned} R_{\hat{\sigma}^2, x_0} &= f(\mathbf{M}_0, \mathbf{M}_0) + 2c_1 tr(\mathbf{M}_0) - 2c_2 f(\mathbf{M}_0, \mathbf{M}_1) \\ &\quad + c_1^2 - 2c_1 c_2 tr(\mathbf{M}_1) + c_2^2 f(\mathbf{M}_1, \mathbf{M}_1) \end{aligned}$$

with

$$f(\mathbf{A}, \mathbf{B}) = tr(\mathbf{A}) tr(\mathbf{B}) + 2 tr(\mathbf{AB})$$

$$\mathbf{M}_0 = (\mathbf{R}_2^{-1} r_2 - \mathbf{R}_1^{-1} r_1)(r_2^t \mathbf{R}_2^{-1} - r_1^t \mathbf{R}_1^{-1}) \mathbf{R}_1$$

$$\mathbf{M}_1 = \mathbf{M} \mathbf{R}_1$$

$$c_1 = 1 - r_1^t \mathbf{R}_1^{-1} r_1$$

$$c_2 = 1 - r_2^t \mathbf{R}_2^{-1} r_2$$

CV and ML estimation

- ML estimation :

$$\hat{\sigma}^2_{ML} = \frac{1}{n} y^t \mathbf{R}_2^{-1} y$$

$var(\hat{\sigma}^2_{ML})$ reaches the Cramer-Rao bound $\frac{2}{n}$

- CV estimation :

$$\hat{\sigma}^2_{CV} = \frac{1}{n} y^t \mathbf{R}_2^{-1} \left[diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-1} y$$

$var(\hat{\sigma}^2_{CV})$ can reach 2

- When $K_2 = K_1$, ML is best. Numerical study when $K_2 \neq K_1$

Criteria for numerical studies (1/2)

Risk on Target Ratio (RTR),

$$RTR(x_0) = \frac{\sqrt{R_{\hat{\sigma}^2, x_0}}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]} = \frac{\sqrt{\mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_0}^2\right)^2\right]}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Bias-variance decomposition

$$R_{\hat{\sigma}^2, x_0} = \left(\underbrace{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c_{x_0}^2\right)}_{\text{bias}}\right)^2 \underbrace{+ var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}_{\text{variance}}$$

Bias on Target Ratio (BTR) criterion

$$BTR(x_0) = \frac{\left|\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c_{x_0}^2\right)\right|}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Criteria for numerical studies (2/2)

$$\Big(\underbrace{RTR}_{\text{relative error}}\Big)^2 = \Big(\underbrace{BTR}_{\text{relative bias}}\Big)^2 + \underbrace{\frac{var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2|y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]^2}}_{\text{relative variance}}$$

Integrated criteria on the prediction domain $\mathcal{X}$

$$IRTR = \sqrt{\int_{\mathcal{X}} RTR^2(x_0) d\mu(x_0)}$$

and

$$IBTR = \sqrt{\int_{\mathcal{X}} BTR^2(x_0) d\mu(x_0)}$$

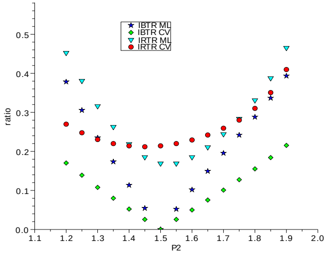
Numerical results

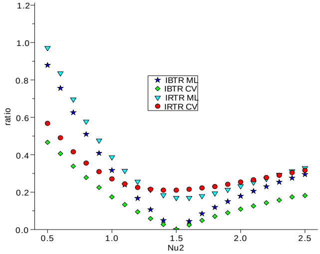
70 observations on $[0,1]^5$. Mean over LHS-Maximin DoE's.

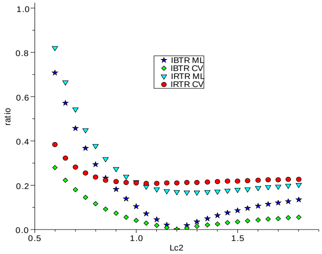
Top : K_1 and K_2 are power-exponential, with $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, and p_2 varying.

Bot left : K_1 and K_2 are Matérn (non-tensorized), with $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.

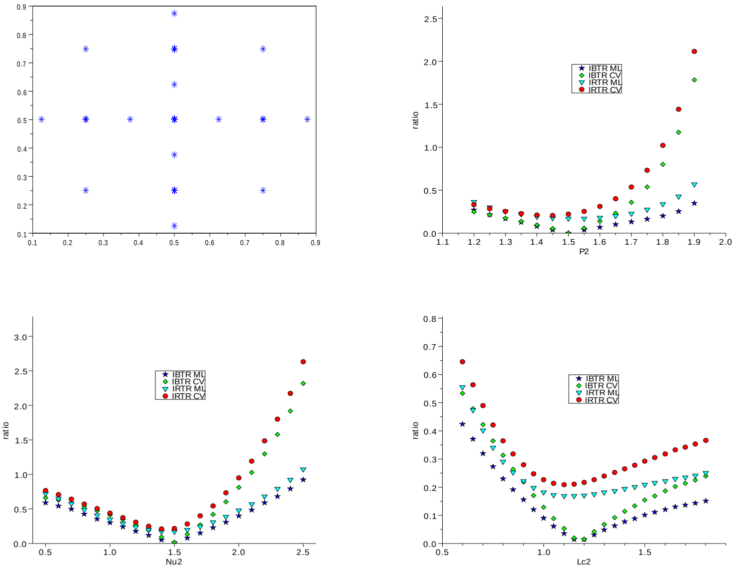
Bot right : K_1 and K_2 are Matérn $\frac{3}{2}$ (non-tensorized), with $l_{c,1} = 1.2$, and $l_{c,2}$ varying.







Case of a regular grid (Smolyak construction)



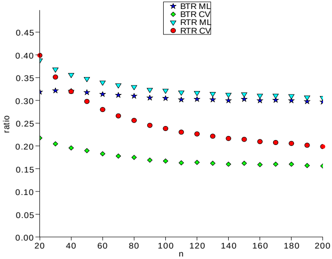
Influence of the number of points

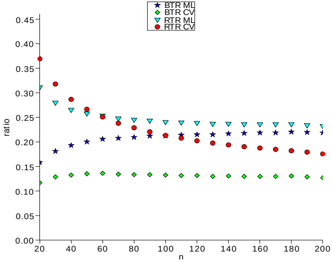
n observations on $[0,1]^5$. Pointwise prediction (center).

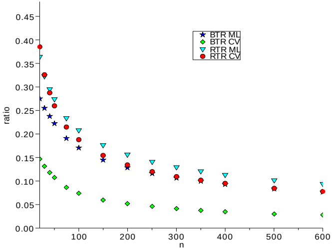
Top : K_1 and K_2 are power-exponential, with $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, and $p_2 = 1.7$.

Bot left : K_1 and K_2 are Matérn (non-tensorized), with $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, and $\nu_2 = 1.8$.

Bot right : K_1 and K_2 are Matérn $\frac{3}{2}$ (non-tensorized), with $l_{c,1} = 1.2$, and $l_{c,2} = 1.8$.







Work on analytical functions

Consider a deterministic function f on $[0,1]^d$

- Ishigami function :

$$f(x_1, x_2, x_3) = \sin(-\pi+2\pi x_1)+7\sin((-\pi+2\pi x_2))^2+0.1\sin(-\pi+2\pi x_1).(-\pi+2\pi x_3)^4$$

- Morris function :

$$f(x) = \sum_{i=1}^{10} w_i(x) + \sum_{1\leq i<j\leq 6} w_i(x)w_j(x) + \sum_{1\leq i<j<k\leq 5} w_i(x)w_j(x)w_k(x) + \sum_{1\leq i<j<k<l\leq 4} w_i(x)w_j(x)w_k(x)w_l(x),$$

with $$w_i(x) = \begin{cases} 2\left(\frac{1.1x_i}{x_i+0.1} - 0.5\right), & \text{if } i = 3,5,7 \\ 2(x_i - 0.5) & \text{otherwise} \end{cases}$$

Comparison criteria

Learning sample $y_{a,1}, \dots, y_{a,n}$. Test sample $y_{t,1}, \dots, y_{t,n_t}$

Mean Square Error (MSE) criterion :

$$MSE = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t,i} - \hat{y}_{t,i}(y_a))^2$$

Predictive Variance Adequation (PVA) criterion :

$$PVA = \left| \log \left(\frac{1}{n_t} \sum_{i=1}^{n_t} \frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\hat{\sigma}^2 c_{t,i}^2(y_a)} \right) \right|$$

We average MSE and PVA over $n_p = 100$ LHS Maximin DoE's. For each DoE : covariance estimation and Kriging prediction

Results with enforced correlation

We use tensorized Exponential and Gaussian correlation functions for the Ishigami function

Correlation model	Enforced hyper-parameters	MSE	PVA
Exponential	$[1, 1, 1]$	2.01	ML : 0.50 CV : 0.20
Exponential	$[1.3, 1.3, 1.3]$	1.94	ML : 0.46 CV : 0.23
Exponential	$[1.20, 5.03, 2.60]$	1.70	ML : 0.54 CV : 0.19
Gaussian	$[0.5, 0.5, 0.5]$	4.19	ML : 0.98 CV : 0.35
Gaussian	$[0.31, 0.31, 0.31]$	2.03	ML : 0.16 CV : 0.23
Gaussian	$[0.38, 0.32, 0.42]$	1.32	ML : 0.28 CV : 0.29

- Misspecified cases : Exponential and Gaussian isotropic
- ML have the highest PVA in the worst misspecification cases

Setting for estimated correlation

- Work on three correlation families
  - Exponential tensorized
  - Gaussian
  - Matérn with estimated regularity parameter
- Work in the isotropic and anisotropic case
  - Case2.i : A common correlation length is estimated
  - Case2.a : d different correlation lengths are estimated

Results for estimated correlation : Ishigami

Function	Correlation model	MSE	PVA
Ishigami	exponential case 2.i	ML : 1.99 CV : 1.97	ML : 0.35 CV : 0.23
Ishigami	exponential case 2.a	ML : 2.01 CV : 1.77	ML : 0.36 CV : 0.24
Ishigami	Gaussian case 2.i	ML : 2.06 CV : 2.11	ML : 0.18 CV : 0.22
Ishigami	Gaussian case 2.a	ML : 1.50 CV : 1.53	ML : 0.53 CV : 0.50
Ishigami	Matérn case 2.i	ML : 2.19 CV : 2.29	ML : 0.18 CV : 0.23
Ishigami	Matérn case 2.a	ML : 1.69 CV : 1.67	ML : 0.38 CV : 0.41

- Gaussian and Matérn are more adapted than exponential because of smoothness ($\rightarrow$ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the exponential case, CV has smaller PVA and smaller or equal MSE
- In the Gaussian and Matérn cases, ML has MSE and PVA slightly smaller

Results for estimated correlation : Morris

Function	Correlation model	MSE	PVA
Morris	exponential case 2.i	ML : 3.07 CV : 2.99	ML : 0.31 CV : 0.24
Morris	exponential case 2.a	ML : 2.03 CV : 1.99	ML : 0.29 CV : 0.21
Morris	Gaussian case 2.i	ML : 1.33 CV : 1.36	ML : 0.26 CV : 0.26
Morris	Gaussian case 2.a	ML : 0.86 CV : 1.21	ML : 0.79 CV : 1.56
Morris	Matérn case 2.i	ML : 1.26 CV : 1.28	ML : 0.24 CV : 0.25
Morris	Matérn case 2.a	ML : 0.75 CV : 1.06	ML : 0.65 CV : 1.43

- Gaussian and Matérn are more adapted than exponential because of smoothness ($\rightarrow$ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the Exponential case, CV has slightly smaller MSE and smaller PVA
- For Gaussian and Matérn 2.a, ML has smaller MSE and PVA
- For Gaussian and Matérn, going from 2.a to 2.i causes much more harm to ML than CV

Conclusion

Conclusion

- We study robustness relatively to prediction mean square errors and point-wise mean square error estimation
- For the variance estimation, CV is more robust than ML to correlation function misspecification
- This is not true for the Smolyak construction we tested
- In the general case of correlation function estimation $\rightarrow$ this is globally confirmed in a case study on analytical functions

Possible perspectives

- Quantify the incompatibility of a DoE for CV ?
- Problem of the choice of the CV procedure

Current work :

- In an expansion asymptotic context, is the regular grid a local optimum for covariance function estimation ?
- Work on ML and CV estimators